BIOGRAPHICAL SKETCH

Provide the following information for the Senior/key personnel and other significant contributors.
Follow this format for each person. DO NOT EXCEED FIVE PAGES.

NAME: Lu, Zhiyong

eRA COMMONS USER NAME (credential, e.g., agency login): zhiyonglu

POSITION TITLE: Senior Investigator

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable. Add/delete rows as necessary.)*

INSTITUTION AND LOCATION	DEGREE (if applicable)	END DATE MM/YYYY	FIELD OF STUDY
Nanjing University	B.S.	07/2001	Computer Science
University of Alberta	M.S.	07/2003	Computer Science
University of Colorado School of Medicine	Ph.D.	07/2007	Bioinformatics

A. Personal Statement

As an NIH Senior Investigator, I have devoted my career largely to research toward biomedical text mining, natural language processing (NLP), machine learning, and related areas. My long-term research goal is to develop computational methods to better understand the natural language in biomedical text in order to accelerate knowledge discovery and improve health. A key to reaching this goal is to be able to accurately capture such biological entities as genes, drugs and diseases and their relationships from free text in an automated fashion. Despite many attempts by others in the past, biomedical named entity recognition and information extraction remains a challenging task. Thus, one thread of my research is to build novel machine-learning based methods (e.g. **PubTator**, **DNorm**) for advancing the state of the art and to explore their real-world uses (e.g. Assisting Biocuration). Moreover, with colleagues from both inside and outside of the NIH, I also lead community-wide efforts (e.g. **BioCreative**) to build resources and to organize worldwide challenge events for such problems. Another unique aspect of my work is that I lead the overall efforts to improve search quality and usability in NCBI production literature resources, in my role as its Deputy Director for Literature Search. Several of my research has been successfully integrated and used by millions of users in PubMed. More recently, I have directed the development of PubMed's new relevance search algorithm called **Best Match** and the new **PubMed Labs**. In addition to the biomedical literature, our research has also been successfully applied to other medical data such as electronic medical records (EMRs) – e.g. a recent effort resulted in the release of **ChestX-ray8**: one of the largest publicly available chest x-ray datasets to the scientific community – as well as medical images, e.g. we recently developed a novel data-driven approach (**DeepSeeNet**) for autonomous age-related macular degeneration (AMD) diagnosis with its performance exceeding human ophthalmologists using deep learning. Since 2004, I have co-authored over 160 peer-reviewed journal, review, and conference proceedings articles.

1. Wei CH, Kao HY, Lu Z. PubTator: a web-based text mining tool for assisting biocuration. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W518-22. PubMed PMID: 23703206; PubMed Central PMCID: PMC3692066.
2. Leaman R, Islamaj Dogan R, Lu Z. DNorm: disease name normalization with pairwise learning to rank. Bioinformatics. 2013 Nov 15;29(22):2909-17. PubMed PMID: 23969135; PubMed Central PMCID: PMC3810844.
3. Fiorini N, Canese K, Bryzgunov R, Radetska I, Gindulyte A, Latterner M, Miller V, Osipov M, Kholodov M, Starchenko G, Kireev E, Lu Z. Best Match: New relevance search for PubMed. PLoS Biol. 2018;16(8):e2005343. PubMed PMID: 30153250; PubMed Central PMCID: PMC6112631
4. Fiorini N, Leaman R, Lipman DJ, Lu Z. How user intelligence is improving PubMed. Nat Biotechnol. 2018;10.1038/nbt.4267. doi:10.1038/nbt.4267. PubMed PMID: 30272675;

B. Positions and Honors

Positions and Employment

2007 – 2009	Staff Scientist, National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), National Institutes of Health (NIH)
2009 – 2011	Associate Investigator, National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), National Institutes of Health (NIH)
2011 – 2016	Earl Stadtman Investigator (tenure-track), National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), National Institutes of Health (NIH)
2016 –	Senior Investigator (early tenure), National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), National Institutes of Health (NIH)
2017 –	Deputy Director for Literature Search, National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), National Institutes of Health (NIH)

Other Experience and Professional Memberships

2004 –	Member, International Society for Computational Biology (ISCB)
2009 –	Member, BioCreative (www.biocreative.org) Organizing Committee
2011 –	Referee, Grant applications from National Institutes of Health (NIH), National Science Foundation (NSF), National Sciences and Engineering Research Council of Canada (NSERC), UK Royal Society Industry Fellowships, The Netherlands Organization for Scientific Research, UK Medical Research Council (MRC)
2012 –	Member, *DATABASE* (Oxford University Press) Editorial Board
2012 –	Member, BioASQ Scientific Advisory Board (European Commission funded)
2013 –	Associate Editor, *BMC Bioinformatics*
2013 –	Member, International Society for Biocuration (ISB)
2013	Chair, IEEE International Conference on Health Informatics (ICHI) Program Committee
2013 –	Member, IEEE International Conference on Health Informatics Steering Committee
2013 – 2015	Session Chair, Pacific Symposium on Biocomputing (PSB)
2015	Member, NLM/NIH Lindberg/King Lectureship Committee
2015	Member, International Society for Biocuration (ISB) Executive Nominating Committee
2015 –	Member, Biomedical Linked Annotation Hackathon Organizing Committee
2015 – 2016	F1000 Faculty Member
2015	Special Issue Editor, *BioMed Research International*
2016	External Referee, National Central University Faculty promotion (Taiwan)
2016 –	Member, *Journal of Healthcare Informatics Research* Editorial Board
2016	Member, ACL - BioNLP Executive Nominating and Election Committee
2016	PC Area Chair, Intelligent Systems for Molecular Biology (ISMB 2016)
2017 – 2018	Member, NIEHS Search Committee for Director, Office of Data Science

Honors

2008	AMIA 2008 Annual Symposium Outstanding Paper Submission
2009	Individual Special Service Award, National Library of Medicine
2010	Special Service Award, National Library of Medicine
2011	Individual Special Service Award, National Library of Medicine
2013	Challenge Award, BioASQ 2013 Global Challenge (http://www.bioasq.org/)
2013	Top Performance, Disease normalization task, CLEF eHealth Global Challenge
2013	Top Performance, Chemical entity mention task, BioCreative IV Global Challenge
2014	Individual Special Service Award, National Library of Medicine
2014	Special Group Service Award, National Library of Medicine
2014	Trans-insight Awards for Semantic Intelligence
2015	Top Performance, CHEMDNER-patents task, BioCreative V Global Challenge

2017	Special Group Service Award, National Library of Medicine
2017	ChemProt-Elsevier Award
2017	NIH Clinical Center Director's Award
2018	Highly Cited Researcher, Clarivate Analytics
2018	Special Group Service Award, National Library of Medicine

C. Contribution to Science

1. Improving PubMed Search for the Biomedical Literature

We have successfully investigated PubMed user's information needs and search behaviors through large-scale search log analysis over the years. Such investigations not only advance our knowledge about biomedical literature search but also have resulted in measurable improvements in PubMed. Based on the results of this study, we have developed several PubMed search assistants such as Automatic Query Suggestion, Author Name Disambiguation, and recently a new relevance search algorithm called **Best Match**. Taken together, these computational features are being used by **millions of PubMed users each day**. More recently, we lead the development of **PubMed Labs** (http://www.pubmed.gov/labs), a new test site towards PubMed 2.0, for experimenting new ways for improve the search quality and usability for the biomedical literature and collecting user feedback.

a. Islamaj Dogan R, Murray GC, Névéol A, Lu Z. Understanding PubMed user search behavior through log analysis. Database (Oxford). 2009;2009:bap018. PubMed PMID: 20157491; PubMed Central PMCID: PMC2797455.
b. Fiorini N, Lipman DJ, Lu Z. Towards PubMed 2.0. Elife. 2017 Oct 30;6. PubMed PMID: 29083299; PubMed Central PMCID: PMC5662282.
c. Fiorini N, Canese K, Bryzgunov R, Radetska I, Gindulyte A, Latterner M, Miller V, Osipov M, Kholodov M, Starchenko G, Kireev E, Lu Z. Best Match: New relevance search for PubMed. PLoS Biol. 2018;16(8):e2005343. PubMed PMID: 30153250; PubMed Central PMCID: PMC6112631
d. Fiorini N, Leaman R, Lipman DJ, Lu Z. How user intelligence is improving PubMed. Nat Biotechnol. 2018;10.1038/nbt.4267. doi:10.1038/nbt.4267. PubMed PMID: 30272675

2. Innovative Text Mining Algorithm Development via Machine Learning

Our group has successfully developed several novel computational methods and natural language processing (NLP) tools for extracting facts and data automatically from plain text. Our research advances the state of the art of text-mining research in information extraction: Not only are our proposed methods highly innovative (e.g. **DNorm**: the first machine-learning algorithm for disease name normalization), they have also been shown to achieve the **highest performance in six global competition-like challenge tasks** held from 2013 to 2018 (e.g. **tmChem**: two-times top-performing method for chemical name recognition in BioCreative IV & V Challenges).

a. Leaman R, Islamaj Dogan R, Lu Z. DNorm: disease name normalization with pairwise learning to rank. Bioinformatics. 2013 Nov 15;29(22):2909-17. PubMed PMID: 23969135; PubMed Central PMCID: PMC3810844.
b. Leaman R, Wei CH, Lu Z. tmChem: a high performance approach for chemical named entity recognition and normalization. J Cheminform. 2015 Jan 19;7(Suppl 1 Text mining for chemistry and the CHEMDNER track):S3. PubMed PMID: 25810774; PubMed Central PMCID: PMC4331693.
c. Wei CH, Phan L, Feltz J, Maiti R, Hefferon T, Lu Z. tmVar 2.0: Integrating genomic variant information from literature with dbSNP and ClinVar for precision medicine. Bioinformatics. 2017 Sep 1. PubMed PMID: 23969135.
d. Singhal A, Simmons M, Lu Z. Text Mining Genotype-Phenotype Relationships from Biomedical Literature for Database Curation and Precision Medicine. PLoS Comput Biol. 2016 Nov 30;12(11):e1005017. PubMed PMID: 27902695; PubMed Central PMCID: PMC5130168.

3. Widely-used Text Mining Services

A major contribution of text mining or bioinformatics research is to develop software tools for the scientific community. Thus, many of our text-mining algorithms have been implemented as open-source software tools with thousands of downloads since 2013. Furthermore, we developed large-scale text-mining web services such as **PubTator** (https://www.ncbi.nlm.nih.gov/research/bionlp/pubtator) to provide literature annotation and on-demand annotation of arbitrary text in real time, which allows those non-NLP specialists to use text mining in their work without a significant investment in infrastructure or methodology. Since April 2015, our web services have been used **~300 million** times by researchers from all around the world in various scenarios (e.g. assisting Biocuration at scale).

a. Wei CH, Kao HY, Lu Z. PubTator: a web-based text mining tool for assisting biocuration. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W518-22. PubMed PMID: 23703206; PubMed Central PMCID: PMC3692066.
b. Wei CH, Leaman R, Lu Z. Beyond accuracy: creating interoperable and scalable text-mining web services. Bioinformatics. 2016 Jun 15;32(12):1907-10. PubMed PMID: 26883486; PubMed Central PMCID: PMC4908316.
c. Leaman R, Lu Z. TaggerOne: joint named entity recognition and normalization with semi-Markov Models. Bioinformatics. 2016 Sep 15;32(18):2839-46. PubMed PMID: 27283952; PubMed Central PMCID: PMC5018376.
d. Poux S, Arighi CN, Magrane M, Bateman A, Wei CH, Lu Z, Boutet E, Bye-A-Jee H, Famiglietti ML, Roechert B, UniProt Consortium T. On expert curation and scalability: UniProtKB/Swiss-Prot as a case study. Bioinformatics. 2017 Nov 1;33(21):3454-3460. PubMed PMID: 29036270.

4. Machine Learning for Healthcare

Mining EMRs and medical images has the potential to lead to improvement in patient care as such data contain rich information for large patient populations. We have recently text-mined over **100,000 radiology reports** where our algorithm generated "weak" training labels to enable the development of advanced deep learning methods for automatically reading and classifying chest X-ray images. This work has also resulted in the release of **ChestX-ray8** (https://nihcc.app.box.com/v/ChestXray-NIHCC): one of the largest publicly available chest x-ray datasets to the scientific community. We have also conducted research to assist in the screening of age-related macular degeneration (AMD): a leading cause of vision loss in Americans 60 and older. By leveraging cutting-edge deep learning techniques and repurposing "big" imaging data from a major AMD clinical trial, we developed a novel data-driven approach (**DeepSeeNet**) for autonomous AMD diagnosis with its performance exceeding human ophthalmologists (retinal specialists in this case). Such a result highlights the potential of deep learning systems to assist early disease detection and enhance clinical decision-making processes.

a. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Lu Z, Summers R. ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. Proceedings of 2017 IEEE Computer Vision and Pattern Recognition. 2017;2097-2106. https://arxiv.org/abs/1705.02315
b. Peng Y, Dharssi S, Chen Q, Keenan TD, Agron E, Wong WT, Chew EY, Lu Z. DeepSeeNet: A deep learning model for automated classification of patient-based age-related macular degeneration severity from color fundus photographs. Ophthalmology. 2018;S0161-6420(18)32185-7. PubMed PMID: 30471319
c. Wang X, Peng Y, Lu L, Lu Z, Summers RM. TieNet: Text-Image Embedding Network for Common Thorax Disease Classification and Reporting in Chest X-rays Proceedings of 2018 EEE Computer Vision and Pattern Recognition (CVPR), 2018. https://arxiv.org/abs/1801.04334
d. Peng Y, Wang X, Lu L, Bagheri M, Summers R, Lu Z. NegBio: a high-performance tool for negation and uncertainty detection in radiology reports. In AMIA 2018 Informatics Summit. https://arxiv.org/abs/1712.05898

5. BioCreative: Critical Assessment of Information Extraction systems in Biology

To advance the field of text mining, we have also been leading efforts in the area of benchmarking and evaluating BioNLP systems. I am one of the organizing committee members of BioCreative (http://www.biocreative.org), the **first and longest-running community-wide effort** for assessing text mining and information extraction systems applied to the biological domain since 2003. In recent years, our laboratory has successfully co-organized five "competitive" text-mining tasks through the BioCreative forum for addressing some of the most important and challenging issues in the field of text mining research: namely gene name recognition, text mining for biocuration workflow, automatic Gene Ontology annotation, extraction of chemical-disease relations, and protein interactions affected by mutation. In each case, we led the text-mining research community towards some novel and challenging problems that the resolution of which are highly important.

a. Lu Z, Hirschman L. Biocuration workflows and text mining: overview of the BioCreative 2012 Workshop Track II. Database (Oxford). 2012 Nov 17;2012. pii: bas043. PubMed PMID: 23160416; PubMed Central PMCID: PMC3500522.
b. Huang CC, Lu Z. Community challenges in biomedical text mining over 10 years: success, failure and the future. Brief Bioinform. 2016 Jan;17(1):132-44. PubMed PMID: 25935162; PubMed Central PMCID: PMC4719069.
c. Wei CH, Peng Y, Leaman R, Davis AP, Mattingly CJ, Li J, Wiegers TC, Lu Z. Assessing the state of the art in biomedical relation extraction: overview of the BioCreative V chemical-disease relation (CDR) task. Database (Oxford). 2016 Mar 19;2016. pii: baw032. PubMed PMID: 26994911; PubMed Central PMCID: PMC4799720.
d. Arighi CN, Wu CH, Cohen KB, Hirschman L, Krallinger M, Valencia A, Lu Z, Wilbur JW, Wiegers TC. BioCreative-IV virtual issue. Database (Oxford). 2014 May 22;2014. pii: bau039. PubMed PMID: 24852177; PubMed Central PMCID: PMC4030502.

Complete List of Published Work: https://www.ncbi.nlm.nih.gov/research/bionlp/Publications

D. Additional Information: Research Support and/or Scholastic Performance

Ongoing Intramural Activities

NIH/NLM ZIA LM000001 Lu (PI) 2010-Present
Query Log Analysis for Improving User Access to NCBI Web Services
The objective is to improve user access to various molecular biology data including the biomedical literature hosted by NCBI through the analysis of query logs. (FY18: $1,691,145)

NIH/NLM ZIA LM000002 Lu (PI) 2010-Present
Named Entity Recognition and Relationship Extraction in Biomedicine
The objective is to improve the state of the art in automatic identification of various biological entities and their relationships in biomedical text such as biomedical literature and clinical text. (FY18: $2,254,860)